When should stream water be sampled to be most informative for event-based, multi-criteria model calibration?

Wang, Ling ; van Meerveld, H J ; Seibert, Jan

Abstract: Isotope data from streamflow samples taken during rainfall or snowmelt events can be useful for model calibration, particularly to improve model consistency and to reduce parameter uncertainty. To reduce the costs associated with stream water sampling, it is important to choose sampling times with a high information content. We used the Birkenes model and synthetic rainfall, streamflow and isotope data to explore how many samples are needed to obtain a certain model fit and which sampling times are most informative for model calibration. Our results for nine model parameterizations and three events, representing different streamflow behaviours (e.g., fast and slow response, with and without overflow), show that the simulation performance of models calibrated with isotope data from two selected samples was comparable to simulations based on isotope data for all 100 time steps. Generally, samples taken on the falling limb were most informative for model calibration, although the exact timing of the most informative samples was dependent on the runoff response. Samples taken on the rising limb and at peakflow were less informative than expected. These model results highlight the value of a limited number of stream water samples and provide guidance for cost-effective event-based sampling strategies for model calibration.

DOI: https://doi.org/10.2166/nh.2017.197

When should stream water be sampled to be most informative for event-based, multi-criteria model calibration?

L. Wang, H. J. van Meerveld and J. Seibert

ABSTRACT

Isotope data from streamflow samples taken during rainfall or snowmelt events can be useful for model calibration, particularly to improve model consistency and to reduce parameter uncertainty. To reduce the costs associated with stream water sampling, it is important to choose sampling times with a high information content. We used the Birkenes model and synthetic rainfall, streamflow and isotope data to explore how many samples are needed to obtain a certain model fit and which sampling times are most informative for model calibration. Our results for nine model parameterizations and three events, representing different streamflow behaviours (e.g., fast and slow response, with and without overflow), show that the simulation performance of models calibrated with isotope data from two selected samples was comparable to simulations based on isotope data for all 100 time steps. Generally, samples taken on the falling limb were most informative for model calibration, although the exact timing of the most informative samples was dependent on the runoff response. Samples taken on the rising limb and at peakflow were less informative than expected. These model results highlight the value of a limited number of stream water samples and provide guidance for cost-effective event-based sampling strategies for model calibration.

Key words | information content, isotope data, model calibration, sampling frequency, sampling strategy, value of limited data

L. Wang (corresponding author)
H. J. van Meerveld
J. Seibert
Department of Geography,
University of Zurich,
Winterthurerstrasse 190,
CH-8057 Zurich,
Switzerland
E-mail: *ling.wang@geo.uzh.ch*

J. Seibert
Department of Earth Sciences,
Uppsala University,
Sweden

INTRODUCTION

Model parameterization is a long-standing issue in hydrological modelling and has been the focus of many studies and research initiatives, e.g., MOPEX (Model Parameter Estimation Experiment, Duan *et al.* 2006). Multi-criteria model calibration can be used to improve internal model consistency by considering other simulated variables than streamflow. Tracer data (mainly conservative environmental tracers, such as water isotopes and chloride) can be particularly powerful for model calibration because their integrated signal at the catchment scale provides information on runoff sources, flow pathways and water age that cannot be obtained from the discharge data only (Lindström & Rodhe 1986; Kirchner 2003, 2006; Birkel & Soulsby 2015; Hrachowitz *et al.* 2015). For example, McGuire *et al.* (2007) found that model calibration with data from tracer experiments improved parameter identifiability and provided insight into the processes that control hillslope-scale water and solute fluxes. de Grosbois *et al.* (1988) used virtual isotope and streamflow data to calibrate the Birkenes model and showed that the optimized parameter values were always better when both data sets were used for calibration than when only streamflow data was used. Application of

doi: 10.2166/nh.2017.197

the isoWATFLOOD model to several catchments showed that although isotope-based calibration did not necessarily lead to more accurate streamflow simulations, it resulted in a more constrained set of model parameters and, therefore, a more robust model (Stadnyk *et al.* 2013). However, other studies have shown that tracer data do not always help to constrain model parameters (Hooper *et al.* 1988; Seibert *et al.* 2003). One reason is that the model structure has to be changed in order to be able to simulate the tracer data and new parameters have to be added to account for mixing processes (Seibert *et al.* 2003). In addition, several studies have shown that the parameters defining the mixing volumes are less identifiable than the flow parameters. This could be due to a poor performance of the isotope simulations or equifinality of the parameter sets because the isotope data did not contain enough information to identify them (Hooper *et al.* 1988; Page *et al.* 2007; Birkel *et al.* 2010a). For example, Dunn & Bacon (2008) used the STREAM model to simulate the response of isotope and chloride concentrations in streamflow with limited success and attributed this to the uncertainties inherent in the input data and the model (both model structure and parameterization).

While isotope and chemical data can be very useful for model calibration, high resolution time series of such data are not regularly available. On the other hand, several studies have shown the value of limited non-continuous data (at single points in time) for lake water levels (Lindström 2016), streamflow (Perrin *et al.* 2007) and groundwater levels (Juston *et al.* 2009). McIntyre & Wheater (2004) tested the value of limited stream phosphorus data for the calibration of a stream phosphorus model. Their results showed that decreasing the total number of samples in a two-month period from 62 (daily) to four (event-based) samples led to only a slight decrease in model calibration performance, especially when there were errors in phosphorus concentrations and model structure. This was partly caused by the dynamic information content of the data, with low flow data being information-poor and possibly detrimental. Using virtual data, Raat *et al.* (2004) found that sampling every 14 days for nitrate and ammonium concentrations in stream water was the most cost-effective monitoring strategy for the calibration of a nitrogen cycle model. However, other studies have shown that the temporal resolution of tracer data (both precipitation and stream water) significantly affects the model performance. Birkel *et al.* (2010a) used different temporal resolutions for the precipitation input in the CIM model and found that model performance increased when using higher resolution data for the precipitation inputs. Birkel *et al.* (2011) reported that daily stream water sampling may not capture important hydrological and isotopic intra-event dynamics, especially for small catchments. Dunn & Bacon (2008) found that weekly precipitation and stream water samples were insufficient to simulate the overall variability in the isotopic composition of stream water, although the streamflow simulation was acceptable.

The aim of this study was to understand how event-based stream water sampling strategies affect model calibration. We, therefore, used the Birkenes model (Christophersen & Wright 1981; de Grosbois *et al.* 1988; Hooper *et al.* 1988) with synthetic rainfall, streamflow and isotope data to answer the following questions: (1) Do a few isotope samples taken during an event allow calibration of a coupled flow and tracer model? (2) When during an event should stream water isotope samples be taken to be most informative for model calibration?

METHODS

The Birkenes model and the nine selected parameterizations

The Birkenes model is a coupled flow and tracer model (hydrochemical model) that was developed to simulate streamflow and the isotopic composition of stream water in the Birkenes catchment in Norway (Hooper *et al.* 1988). The Birkenes model was selected for this study because: (i) it is a simple model with a limited number of parameters and few requirements for the input data; (ii) it is suitable for event simulation because it was developed to predict short-term changes in hydrochemistry; (iii) its model structure and parameters form the basis for several newer conceptual models that include tracer simulations (Fenicia *et al.* 2008; Birkel *et al.* 2010a; Soulsby *et al.* 2015); and (iv) it is well-known and has been applied to catchments in different countries (Grip *et al.* 1985; Seip *et al.* 1985; de Grosbois

et al. 1986; Rustad et al. 1986; Wheater et al. 1986; Hooper et al. 1988; Neal et al. 1988).

The Birkenes model consists of two linear reservoirs: reservoir A represents a quick response (Q_A), while reservoir B has a slower response (Q_B) (Figure 1). The model has seven parameters: three dimensional parameters (AMIN, BMIN and BSIZE), two rate parameters (AK and BK), one routing parameter (AKSMX) and a constant baseflow (Q_{BASE}) (Hooper et al. 1988). Parameter AMIN represents the threshold storage in reservoir A for quick response flow (Q_A) to occur, while parameter BMIN represents the threshold storage to produce the slow response flow (Q_B) from reservoir B. The sum of BMIN and BSIZE represents the maximum storage in reservoir B. Overflow (Q_{OVER}) occurs when reservoir B is full. The two rate parameters (AK and BK) describe the fluxes out of reservoirs A and B as a function of the storage in the reservoirs. The routing parameter (AKSMX) defines the fraction of water that flows from reservoir A into reservoir B. Parameter QBASE represents the constant baseflow (Q_{BASE}) to the stream (i.e., it is unaffected by the storage in reservoir B) and is usually set to the minimum observed streamflow (Figure 1; de Grosbois et al. 1988). Evaporation from reservoir A (E_A) was set to 0.03 mm h^{-1} and it was assumed that there was no evaporation from reservoir B.

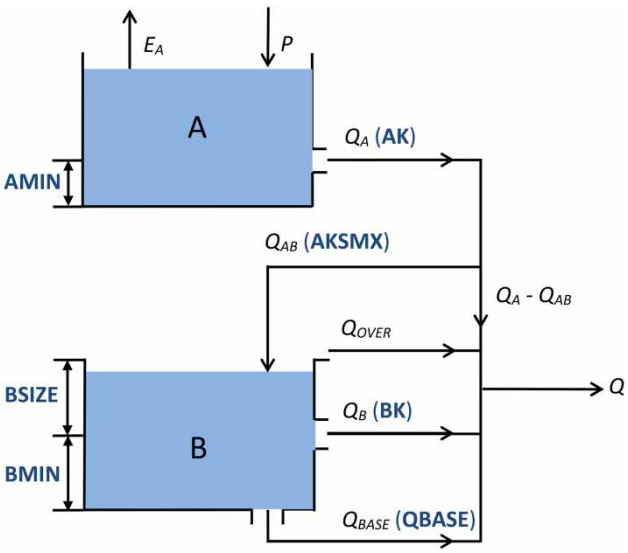

Figure 1 | Schematic diagram of Birkenes model with the model parameters written in bold and state variables in italic (modified after Hooper et al. 1988).

Similar to other coupled flow and tracer models (see Birkel & Soulsby (2015) for a review), our study focuses on conservative tracers. We chose oxygen-18 as the target tracer for model simulation but it could have been deuterium or another conservative tracer as well. The model assumes complete mixing within each of the two reservoirs. The concentration in reservoir B is also assigned to Q_{OVER} (de Grosbois et al. 1988). Isotope fractionation is not included in the model but we expect it to have a small influence on the results when evaporation from the soil and lakes is limited.

Nine different parameterizations of the Birkenes model were used to represent different streamflow behaviours (e.g., fast and slow response, with and without overflow). The first parametrization (P1) is based on the parameter values from Christophersen & Wright (1981) for their manual fit of the model to the observations in the Birkenes catchment. This parameter set was also used by de Grosbois et al. (1988) in their study on multiple signal calibration (based on isotope data and streamflow). For the eight other parameterizations, the values of parameters BSIZE, BK, AK and AKSMX were adjusted to obtain streamflow time series that are dominated by different flow pathways (different amounts of flow from the fast and slow reservoir and overflow) and have different response times (Table 1 and Table S1 and Figure S1).

For each of the nine parameterizations, we simulated streamflow and the isotopic composition of stream water during three rainfall events with a total rainfall of 12 mm (E1), 24 mm (E2) and 48 mm (E3) and a constant rainfall intensity of 4 mm h^{-1} (which is reasonable for the original Birkenes catchment and climate). Initial tests with a rainfall intensity of 8 mm h^{-1} showed only a minor effect of doubling the rainfall intensity on the modelled streamflow and tracer response compared to the effect of doubling the event size. In order to minimize the total number of potential model simulations, we therefore decided to keep the rainfall intensity constant and focus on the effect of event size and the corresponding changes in the amount of fast flow (Q_A), slow flow (Q_B) and overflow (Q_{OVER}).

The model warming up period consisted of 100 weeks with the same event at the start of each week. The isotopic composition of rainfall ($\delta^{18}O$) was set to −10‰ for the first 95 weeks, and to −15‰, −10‰, −5‰, −10‰ and −5‰ for the following 5 weeks to obtain a different initial

Table 1 | Parameter values for the nine parameterizations (P1–P9)

Parameterization	P1	P2	P3	P4	P5	P6	P7	P8	P9
AMIN [mm]	13	13	13	13	13	13	13	13	13
BMIN [mm]	40	40	40	40	40	40	40	40	40
BSIZE [mm]	40	**25**	**15**	40	40	40	40	40	40
AK [h^{-1}]	3.33×10^{-2}	3.33×10^{-2}	3.33×10^{-2}	3.33×10^{-2}	3.33×10^{-2}	3.33×10^{-2}	3.33×10^{-2}	$\mathbf{1.67\times10^{-2}}$	$\mathbf{1.67\times10^{-1}}$
BK [h^{-1}]	1.90×10^{-3}	1.90×10^{-3}	1.90×10^{-3}	$\mathbf{3.80\times10^{-4}}$	$\mathbf{9.50\times10^{-3}}$	1.90×10^{-3}	1.90×10^{-3}	1.90×10^{-3}	1.90×10^{-3}
AKSMX [–]	0.75	0.75	0.75	0.75	0.75	**0.5**	**0.25**	0.75	0.75
QBASE [mm h^{-1}]	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Change in catchment response	Birkenes catchment	Smaller reservoir B	Smallest reservoir B	Reservoir B drains slower	Reservoir B drains faster	Less water flows from reservoir A to reservoir B	Even less water flows from reservoir A to reservoir B	Reservoir A drains slower	Reservoir A drains faster

Values in bold for P2–P9 indicate changes compared to P1. See Figure S1 for the corresponding hydrographs.
The parameter values for P1 are similar to the values for the Birkenes catchment (Christophersen & Wright 1981).

isotopic composition in reservoirs A and B. The isotopic composition of the rainfall during the event of interest (week 101) was set to −15‰.

The simulated streamflow and isotopic composition of stream water for the 27 model simulations (the three different events (E1–E3) for the nine different parameterizations (P1–P9)) were used as observations. We used this synthetic data as our observed time series because: (i) this way we would know the isotopic composition of stream water at every potential time step (hourly in this study), whereas it is difficult to collect such high temporal resolution data in reality; and (ii) it is theoretically possible to obtain a perfect model fit and the model results are, therefore, not affected by any errors in the model or the data that may otherwise affect our interpretation of the time of the most informative samples. In order to evaluate the value of a limited number of stream water samples for model calibration, we pretended that all hourly streamflow data and only a subset of the isotope data ($n = 0,1,2,\ldots n$ samples) were available for model calibration. For model validation we assumed that all isotope data were available ($n = 100$ samples).

Model calibration and parameter optimization

The SCE-UA method (Duan *et al.* 1992, 1993) was chosen for automatic parameter optimization because the algorithm is considered reliable and fast (Francés *et al.* 2007), the Matlab code and guidelines to apply the SCE-UA method are available online (Duan *et al.* 1994), and the method has been tested and implemented in several studies (Yapo *et al.* 1996; Francés *et al.* 2007). For each calibration, 25 seeds were used to account for the influence of the initial selection of the parameter values (Figure 2, Step 1). The parameter ranges were set to 0.2 to 5 times the actual parameter for each parameterization (Table 1), except for AK, which was set to 0.5 to 5 times the actual parameter value because the value of AK should be larger than the value of BK. The model was calibrated by minimizing the combined objective function (Equation (1)):

$$F = \sqrt{\frac{F_Q^2 + F_C^2}{2}} \tag{1}$$

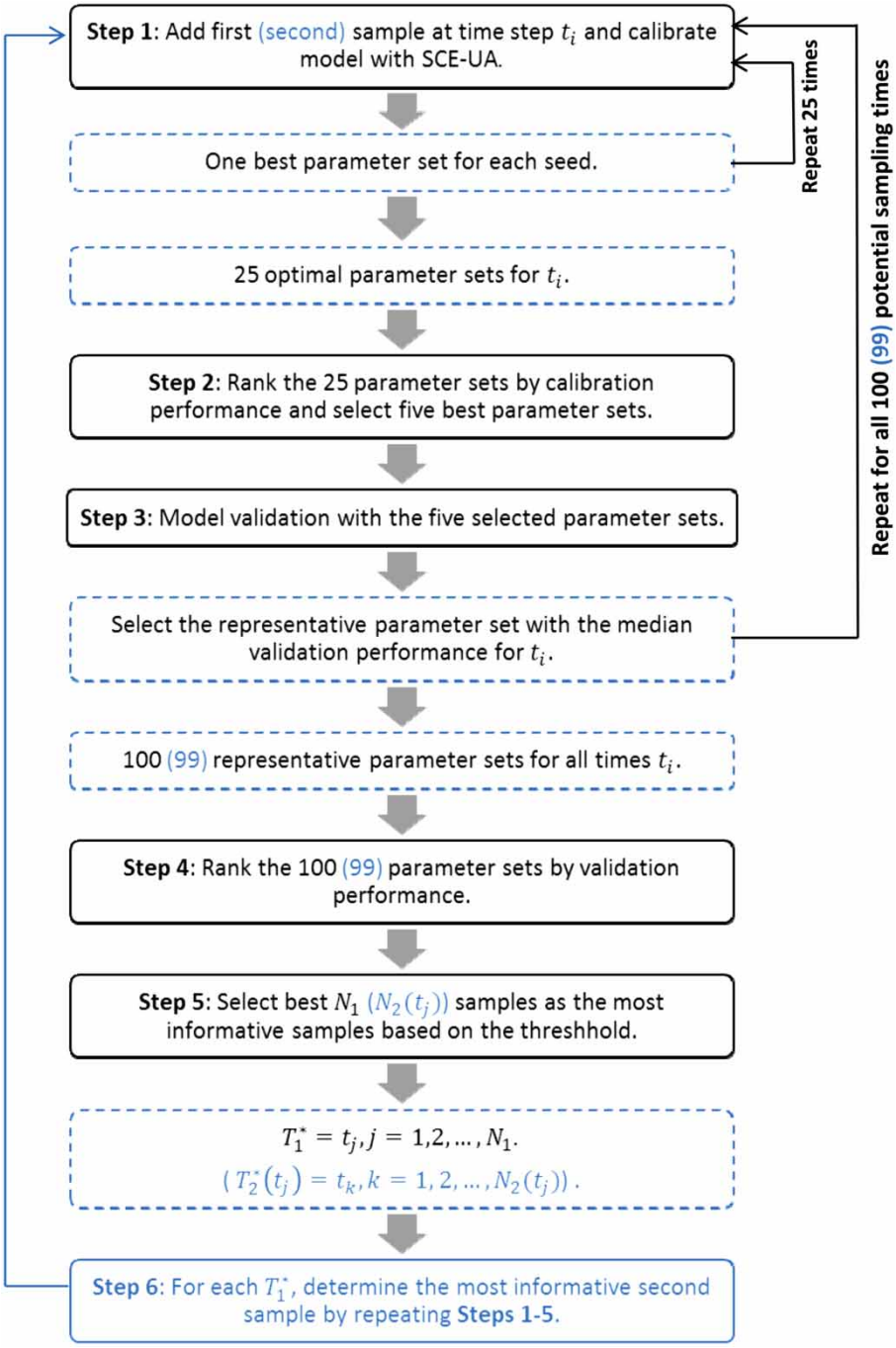

Figure 2 | Flowchart of the model steps to find the most informative first and second samples (T_1^* and T_2^*). Dashed boxes show the results from previous steps. Note that the information in parentheses is used to find the most informative second samples.

where F_Q is the objective function for streamflow (Equation (2)) and F_C is the objective function for the isotopic composition of stream water (Equation (3)), where:

$$F_Q = \frac{1}{m}\sum \frac{|Q_{\mathrm{obs}(i)} - Q_{\mathrm{sim}(i)}|}{Q_{\max} - Q_{\min}} \quad (2)$$

where $Q_{obs(i)}$ is the observed streamflow at time i, $Q_{sim(i)}$ is the simulated streamflow at time i, Q_{max} is the maximum observed streamflow, Q_{min} is the minimum observed streamflow and m is the number of streamflow measurements ($m = 100$).

$$F_C = \frac{1}{n}\sum \frac{|C_{obs(i)} - C_{sim(i)}|}{C_{max} - C_{min}} \quad (3)$$

where $C_{obs(i)}$ is the observed isotopic composition of stream water at time i, $Q_{sim(i)}$ is the simulated isotopic composition of stream water at time i, n is the number of stream water isotope samples (varying between 0 and 100) and C_{max}–C_{min} is the range in the isotopic composition of stream water, which was set to 5‰ (i.e., the difference between the long-term mean isotopic composition of rainfall and the isotopic composition of the rainfall for the event of interest).

The three objective functions (Equations (1)–(3)) vary between 0 and 1, where 0 means a perfect fit and larger values indicate poorer simulations. The combined objective function (Equation (1)) and the normalization of the objective functions for streamflow and the isotopic composition of stream water (Equations (2) and (3)) were chosen to equally weigh the model performance for streamflow and the isotopic composition of stream water and to avoid bias to either of these two.

Model validation and selection of the most informative stream water samples

For each parameterization and event, the model was first calibrated without any information on the isotopic composition of stream water ($n = 0$). The five best models (from the 25 seeds) were validated using all the information on the stream water isotopic composition ($n = 100$ samples). The parameter set with the median value of the combined objective function (Equation (1)) for the validation was chosen as the representative parameter set for the calibration without any information on stream water quality.

Then, the model was calibrated using one measurement of the isotopic composition of stream water (i.e., $n = 1$ sample). For each potential sampling time, the five best calibrations (from the 25 seeds) were again used for validation based on the full data set and the parameter set with the median value of the combined objective function for the validation was selected as the representative simulation for the calibration of the model with the isotope data for that sampling time. This procedure was repeated for all 100 time steps (96 event samples and four pre-event samples) (Figure 2, Steps 1–3). We then ranked the value of the objective function for the validation of the selected models for the 100 sampling times (Figure 2, Step 4) and chose all sampling times with a value of the objective function that was within two times the difference between the third and fifth highest ranked sampling time (T_1*) because the values of the objective function for the validation were not always significantly different for the high-ranked sampling times. This ensured that at least the five best sampling times were chosen ($N_1 \geq 5$) and avoided exclusion of sampling times with an almost equally good validation (Figure 2, Step 5). These selected N_1 best sampling times are regarded as the intelligently selected and most informative sampling times.

For each of the selected most informative sampling times ($T_1^* = t_j$, $j = 1, 2, \ldots, N_1$), this process (Figure 2, Steps 1–5) was repeated by adding a second sample for the calibration for all remaining 99 potential sampling times (Figure 2, Step 6). For each selected most informative first sampling time t_j, we get $N_2(t_j)$ most informative second sampling times $T_2^*(t_j)$. The N_2 combinations of t_j and t_k are considered the most informative sampling pairs.

When the maximum error in the concentration for the validation for the models with the most informative sampling pairs was larger than 0.1‰, this process was repeated to find the most informative third (and fourth) sampling times as well. The 0.1‰ maximum error in stream water isotopic composition was chosen as the cut-off value because it is similar to the sample analytical uncertainty (Leibundgut *et al.* 2009; Stadnyk *et al.* 2013).

Comparison to benchmarks

In order to determine the importance of the sampling time for model calibration, the values of the objective functions for the validation and the maximum error in the concentration of the models calibrated with the selected (i.e., most informative) samples were compared to models calibrated with randomly selected sampling times and sampling times based on the streamflow dynamics.

Random selection

For the models calibrated with only one sample, we used the median of the objective function for the validation and the median maximum error in the isotopic composition of stream water for all potential sampling times as the benchmark (*B-R1*). For the comparison of the models calibrated with two samples, we calibrated the model with 1,000 random pairs of samples that were taken at least 5 hours apart. For each randomly selected sample pair, we used 25 seeds for model calibration, selected the five best seeds, calculated the value of the combined objective function for the validation for these five seeds, and selected the seed with the medium value as the representative model. We used the median values of the objective functions for the validation and the median maximum error in the concentration of these 1,000 randomly selected sampling pairs as the benchmark (*B-R2*). Similarly, for the models calibrated with three (or more) samples, we selected 5,000 random triplets of sampling times that were at least 5 hours apart and chose the median of the objective function for the validation and the median maximum error as the benchmark (*B-R3* or *B-Rn*).

Based on streamflow dynamics

Hydrologists often try to obtain samples on both the rising and falling limbs because they provide different information. Samples taken close to peakflow are often considered informative as well. For the one-sample benchmark, we therefore used the isotopic composition of stream water at peakflow for model calibration and used the values of the objective functions for the validation and the maximum error in the isotopic composition for this calibration as the benchmark (*B-Q1*). For the two-sample comparison, we used the sample taken at peakflow and either the sample taken at the time that streamflow had reached half of the increase between baseflow and peakflow on the rising limb (*B-Q2r*) or on the falling limb (*B-Q2f*) for model calibration. For the models calibrated with three samples, we selected the sample at peakflow and samples at half of the rising limb and falling limb for model calibration and used this as the benchmark (*B-Q3*).

Parameter information content by dynamic identifiability analysis

The information content of parameters AMIN and BMIN was calculated by Dynamic identifiability analysis (DYNIA) (Wagener *et al.* 2003), as implemented in the Sensitivity Analysis For Everybody (SAFE) toolbox (Pianosi *et al.* 2015), to determine during which periods these parameters were most identifiable. These two parameters were chosen because they could not be identified based on streamflow data alone (see results below). For the DYNIA analysis, the model was run with 100,000 randomly selected parameter sets (changing only the AMIN and BMIN parameters and keeping the other parameters at their real value) and the best 1,000 simulations (i.e., simulations with the smallest value of the combined objective function, Equation (1)) were chosen for each time step. The parameter range of the selected 1,000 simulations for each time step was divided into 20 equally spaced intervals. The information content of the parameter was then calculated for each time step as one minus the relative number of intervals over which the 1,000 selected parameter values were distributed (i.e., if the parameter values of the 1,000 best simulations for that time step were all located in one interval, the information content equalled 0.95, whereas if the parameter values were distributed over all intervals, the information content would be zero).

RESULTS

Rainfall–runoff response classification

The combination of nine model parameterizations (P1–P9) and three rainfall events (E1–E3) resulted in 27 different rainfall–runoff responses (Figure S1) that could be classified into three dominant types. Events for which Q_{BASE} and Q_B contributed more than 80% of total streamflow are classified as slow responses (*S*). The other rainfall runoff responses are classified as fast responses with overflow (*F*) or without overflow (*Fo*) depending on whether Q_{OVER} occurred or not (Table S1).

The manuscript will mainly focus on the results of two parameterizations (P1 and P7) because they show these representative behaviours but the results for all other

parameterizations and the plots for all parameterizations are shown in the Supplementary material. Parameterization 1 (P1, the original Birkenes model) was characterized by the slow response for the small event (P1E1), the fast response for the medium event (P1E2) and the fast response with overflow for the large event (P1E3) (Figure 3, left column). For parameterization 7 (with a smaller value for AKSMX compared to the original Birkenes model), the streamflow response during the small event was characterized by the slow response, while the medium and large events were characterized by the fast response without overflow (Figure 3, right column).

Number of samples needed for model calibration

The models calibrated without any isotope data fit the streamflow well. The maximum errors for streamflow were very small (less than 2×10^{-7} mm h^{-1}; Table 2). However, the maximum error in the simulated isotopic composition of stream water was high for all parameterizations and all events (Table 2). The addition of information from a single intelligently selected sample (i.e., taken at the most informative first sampling time) decreased the value of the combined objective function and improved the fit of the isotopic composition of stream water (Figures 4–6) but slightly increased the maximum error in the modelled streamflow (Table 2). However, the increase in the streamflow error was very small compared to the improvement in the simulation of the isotopic composition of stream water (Table 2). The addition of the information from a second intelligently selected sample (i.e., taken at the most informative second sampling time) improved the model fit even further, with the values of the combined objective function and maximum errors in the isotopic composition of stream water being

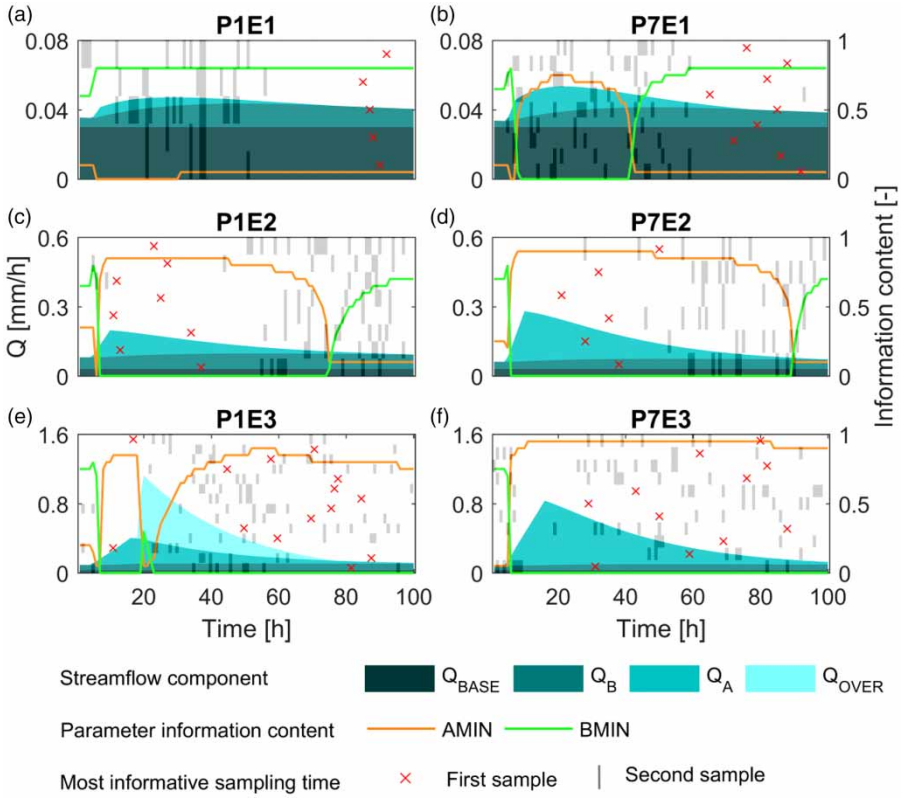

Figure 3 | Streamflow responses for two parameterizations (P1 left and P7 right) during the three rainfall events (small event top row, medium event middle row and large event bottom row), as well as the sampling times of the two most informative streamflow samples, and the information content of a sample with regards to AMIN and BMIN. In each subplot, the most informative first samples are marked with crosses and the most informative second samples that belong to each first sample are marked with grey lines on the same row. The number of selected most informative samples is different for each model parameterization and event because we chose all sampling times with a value of the objective function that was within two times the difference between the value of the combined objective function of the third and fifth ranked sampling time (see text). Note that the y-axis is different for the different events. Please see the online pdf of the paper for the color version of this figure.

Table 2 | Values of the goodness of fit measures for the three different events (E1–E3) for P1 and P7 calibrated without any isotope data ($n = 0$), with one ($n = 1$) or two ($n = 2$) intelligently selected samples and with isotope data for all 100 time steps ($n = 100$)

	n	P1E1	P1E2	P1E3	P7E1	P7E2	P7E3
Max error Q [mm/h]	0	1.13×10^{-9}	2.54×10^{-9}	1.72×10^{-7}	1.15×10^{-9}	1.59×10^{-9}	3.81×10^{-9}
	1	3.46×10^{-9}	3.67×10^{-9}	6.36×10^{-6}	2.39×10^{-9}	9.34×10^{-9}	1.67×10^{-8}
	2	2.82×10^{-9}	3.67×10^{-9}	2.29×10^{-6}	3.25×10^{-9}	6.72×10^{-9}	1.53×10^{-8}
	100	9.86×10^{-9}	5.64×10^{-9}	3.86×10^{-6}	3.40×10^{-9}	2.56×10^{-8}	3.71×10^{-8}
Max error C [‰]	0	4.13×10^{-1}	7.85×10^{-1}	1.44×10^{0}	3.88×10^{-1}	1.24×10^{-0}	1.41×10^{0}
	1	3.69×10^{-2}	3.99×10^{-1}	4.52×10^{-1}	1.04×10^{-1}	2.88×10^{-1}	1.89×10^{-1}
	2	4.71×10^{-7}	2.83×10^{-7}	1.34×10^{-5}	4.37×10^{-7}	4.13×10^{-7}	1.48×10^{-6}
	100	1.74×10^{-6}	4.14×10^{-7}	6.70×10^{-6}	6.45×10^{-7}	8.82×10^{-7}	4.47×10^{-7}
F_Q [–]	0	1.92×10^{-8}	5.84×10^{-9}	1.29×10^{-8}	2.86×10^{-8}	2.44×10^{-9}	1.78×10^{-9}
	1	5.67×10^{-8}	1.20×10^{-8}	4.53×10^{-7}	3.32×10^{-8}	1.12×10^{-8}	6.75×10^{-9}
	2	4.89×10^{-8}	8.86×10^{-9}	1.48×10^{-7}	4.90×10^{-8}	8.75×10^{-9}	6.57×10^{-9}
	100	1.17×10^{-7}	2.24×10^{-8}	9.59×10^{-8}	4.37×10^{-8}	1.32×10^{-8}	2.09×10^{-8}
F_C [–]	0	3.10×10^{-2}	6.35×10^{-2}	9.75×10^{-2}	2.32×10^{-2}	1.23×10^{-1}	1.93×10^{-1}
	1	1.75×10^{-3}	1.17×10^{-2}	1.95×10^{-2}	8.86×10^{-3}	9.24×10^{-3}	2.64×10^{-3}
	2	2.66×10^{-8}	9.40×10^{-9}	3.95×10^{-7}	3.24×10^{-8}	1.33×10^{-8}	2.08×10^{-8}
	100	3.97×10^{-8}	2.09×10^{-8}	6.97×10^{-8}	6.23×10^{-8}	1.09×10^{-8}	1.79×10^{-8}
F [–]	0	2.19×10^{-2}	4.49×10^{-2}	6.89×10^{-2}	1.64×10^{-2}	8.72×10^{-2}	1.37×10^{-1}
	1	1.24×10^{-3}	8.28×10^{-3}	1.38×10^{-2}	6.26×10^{-3}	6.54×10^{-3}	1.87×10^{-3}
	2	4.09×10^{-8}	9.58×10^{-9}	2.98×10^{-7}	4.40×10^{-8}	1.17×10^{-8}	1.62×10^{-8}
	100	8.74×10^{-8}	2.16×10^{-8}	8.38×10^{-8}	5.38×10^{-8}	1.21×10^{-8}	1.94×10^{-8}

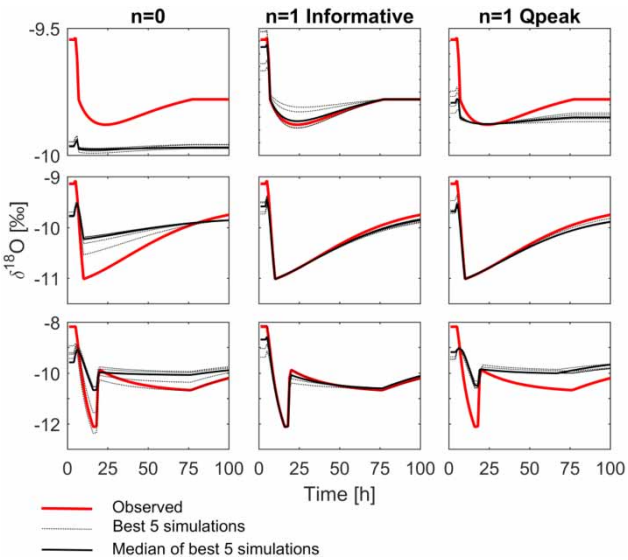

Figure 4 | Observed and modelled time series of the isotopic composition of stream water for the 5 best models calibrated without any isotope data (left column), with the sample taken at the most informative first sampling time (middle column) and the sample taken at peak flow (right column) for model parameterization P1 for the small event (upper row), medium event (middle row) and large event (bottom row). Note that the y-axis is different for the different events.

similar to models calibrated with all isotope data ($n = 100$ samples, see Figures 5 and 6 and Table 2).

Comparison to benchmarks

In general, the models calibrated with one intelligently selected sample (i.e., taken at the most informative first sampling time) resulted in better fits than the one-sample benchmarks (*B-R1* and *B-Q1*) (Figures 5 and 6, and Figures S2 and S3). For event 1 (dominated by the slow response), the median maximum error in the isotopic composition of stream water was less than 0.1‰ when calibrated with the intelligently selected sample for seven of the nine parameterizations. For the benchmark simulations with a randomly selected sample (*B-R1*), this was only the case for two parameterizations (P5E1 and P8E1); for the calibrations with the sample taken at peakflow this was only the case for P5E1. For events 2 and 3, which are dominated by fast response flow (except P5E2 and P5E3, which are dominated by slow response), calibration with one sample resulted in

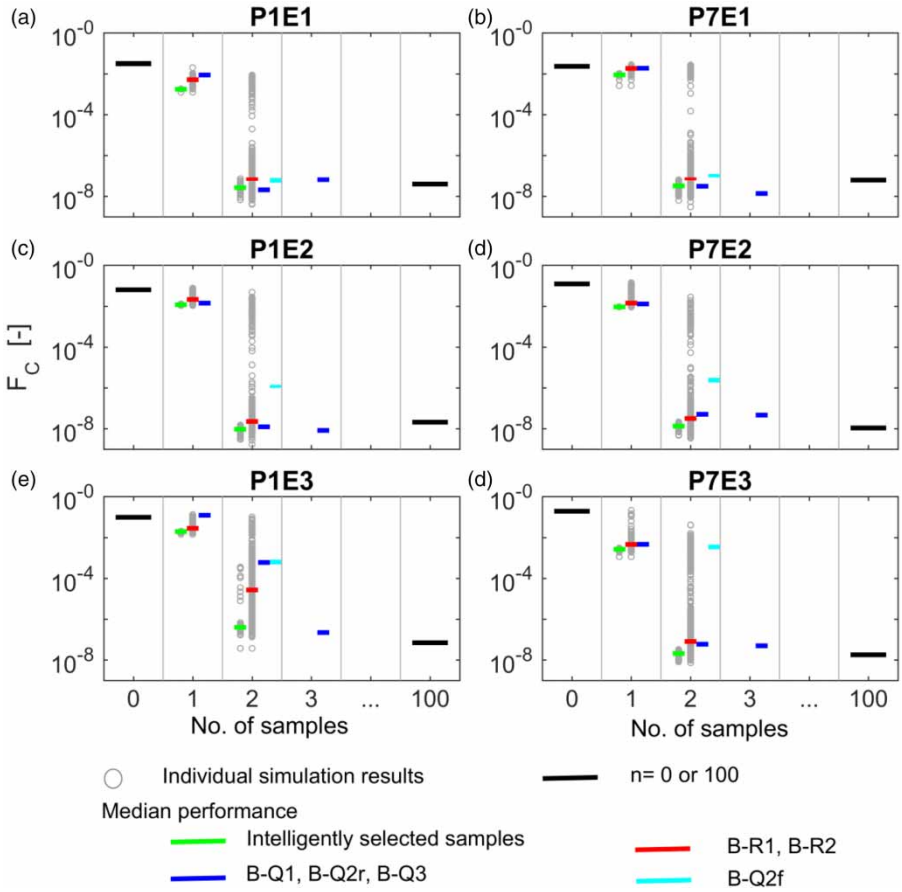

Figure 5 | Comparison of the median value of the objective function for concentration F_C (Equation (3)) for the validation period for two parameterizations (P1 left and P7 right) and three events (small event top row, medium event middle row and large event bottom row) when the model was calibrated without any isotope data ($n = 0$), one sample ($n = 1$), two samples ($n = 2$), three samples ($n = 3$), and all samples ($n = 100$), as well as the values for the individual simulation results (open circles). For one, two and three samples the results are shown for the different sampling strategies (intelligently selected samples and the benchmarks). Please see the online pdf of the paper for the color version of this figure.

maximum errors larger than 0.1‰. For these events, calibration with the intelligently chosen sample performed better than the two benchmarks and had lower median maximum errors in the modelled isotopic composition of stream water, except for P5E2 and P8E2 for which the calibration based on the sample taken at peakflow resulted in a slightly smaller median maximum error (0.87 vs 0.82‰ for P5E2 and 0.43 vs 0.42‰ for P8E2) (Figure S3). These results suggest that when only one sample is available, the timing of the sample influences model calibration and the sample taken at peakflow is generally not the most informative one.

Two intelligently selected samples (i.e., taken at the most informative first and second sampling time) were sufficient to reduce the maximum error in the isotopic composition of stream water below 0.1‰ for all parameterizations and events (Figure 6 and Figure S3). For the calibrations based on two randomly selected samples (*B-R2*), the median value of the maximum error in the isotopic composition of streamflow was also less than 0.1‰, except for P9E1, P5E2, P8E3 and P9E3. However, the range of the values of the maximum error in the isotopic composition was large and for many of the realizations the maximum error was larger than 0.1‰ (Figure 6). For the models calibrated with a sample taken on the midpoint of the rising limb and at peakflow (*B-Q2r*), the maximum errors in the isotopic composition were also smaller than the 0.1‰ threshold, except for P2E2, P3E3, P4E3 and P8E3. Similarly, for the models calibrated with a sample taken on the midpoint of the falling limb and at peakflow (*B-Q2f*), the maximum errors were less than the 0.1‰ threshold for all parameterizations and all events, except for P5E2, P7E3 and P8E3. The good performance of the

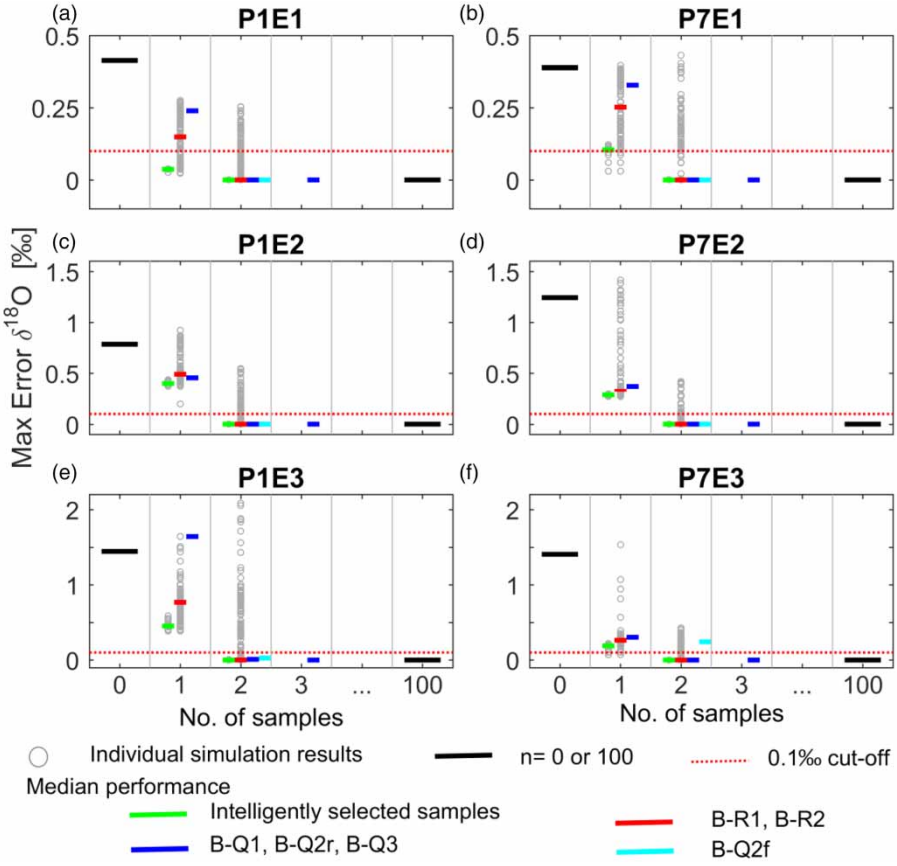

Figure 6 | Comparison of the median maximum error in the simulated isotopic composition of stream water for two parameterizations (P1 and P7) and three events (small event top row, medium event middle row and large event bottom row) when the model was calibrated without any isotope data ($n = 0$), one sample ($n = 1$), two samples ($n = 2$), three samples ($n = 3$), and all samples ($n = 100$), as well as the values for the individual simulation results (open circles). For one, two and three samples the results are shown for the different sampling strategies (intelligently selected samples and the benchmarks). The dashed line indicates the 0.1‰ cut-off for the maximum error. Please see the online pdf of the paper for the color version of this figure.

models calibrated with two samples suggests that for most runoff events (particularly the small and medium events), two samples are sufficient to obtain a good model fit and that the exact sampling time does not matter much, except for the large events for P8 and P9. For P8E3 and P9E3, six and four random samples taken at least 5 hours apart, respectively, were needed to reduce the median maximum error below 0.1‰.

Timing of the most informative samples

First sample

For two-thirds of the 27 streamflow responses, samples taken at the end of the event (between hours 60 and 100) were most informative for model calibration and nearly all (more than 90%) of the most informative first samples were located on the falling limb of the event (crosses in Figure 3 and Figure S1). For the slow response dominated (type *S*) runoff responses, samples taken at the end of the event (between hours 60 and 100) were most informative for model calibration (Figure 3(a) and 3(b)). This corresponds to the time that the fast response (Q_A) had ended. For the fast response dominated runoff responses without overflow (type *F*), a sample taken near or after peakflow was most informative for model calibration (Figure 3(c), 3(d) and 3(f)). For the fast response with overflow (type *Fo*) (Figure 3(e)), a sample taken before overflow starts or when overflow had almost ended but Q_A was still significant was most informative for model calibration.

Second sample

The most informative second sampling time for model calibration was generally several hours before or after the most informative first sample. The exact timing of the second sample did not significantly affect the model results (see the wide area with grey colours in Figure 7 and Figure S4).

Parameter identifiability and information content

Parameters KA, KB and QBASE could be identified based on the calibration with only the streamflow data (Figure 8). Parameters BSIZE and AKSMX could only be identified when overflow occurred (e.g., P1E3, Figure 8(c)). When the slow reservoir (B) was not filled and overflow did not occur, parameters BSIZE and AKSMX could not be identified because their optimized values were linearly correlated and did not affect the simulated streamflow (Figure 8(a) and 8(b)). Parameters AMIN and BMIN (the threshold storage for flow to occur from reservoirs A and B, respectively) could not be identified based on the streamflow data alone. The selected most informative first isotope sample allowed either parameter AMIN or BMIN to be identified. For the slow flow dominated response (type *S*), the most informative first sample allowed determination of parameter BMIN, whereas for the fast flow dominated responses (type *F* and *Fo*), the first sample helped with the determination of parameter AMIN. The addition of a second stream isotope sample allowed the identification of both parameters (Figure 8).

DYNIA was used to understand at what times parameters AMIN and BMIN were most identifiable. The temporal variation in the information content of AMIN and BMIN was mostly opposite: when the information content was high for one parameter, it was low for the other

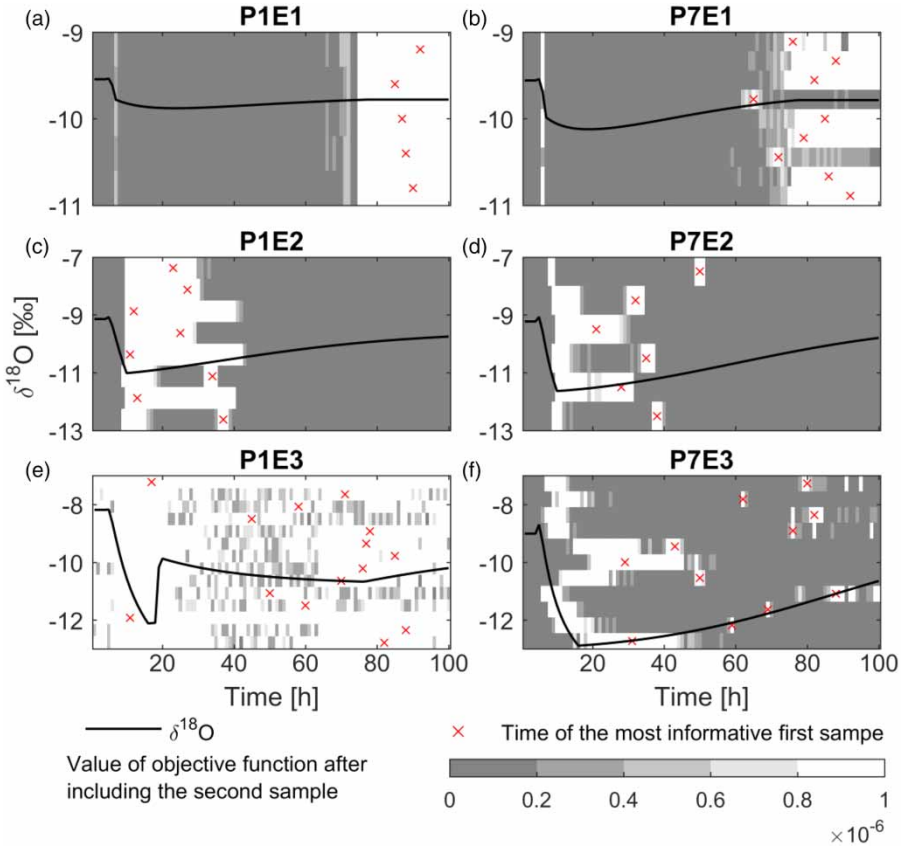

Figure 7 | The observed isotopic composition of stream water for two parameterizations (P1 left and P7 right) and the three different rainfall events (small event top row, medium event middle row, large event bottom row), together with the sampling times of the most informative first samples (crosses) and the values of the combined objective function (Equation (1)) after the second sample has been added for calibration (in grey scale). In each row of each subplot, darker colours represent better model validation results after adding a second sample when the most informative first sample (cross) was fixed. Note that the y-axis is different for the different events.

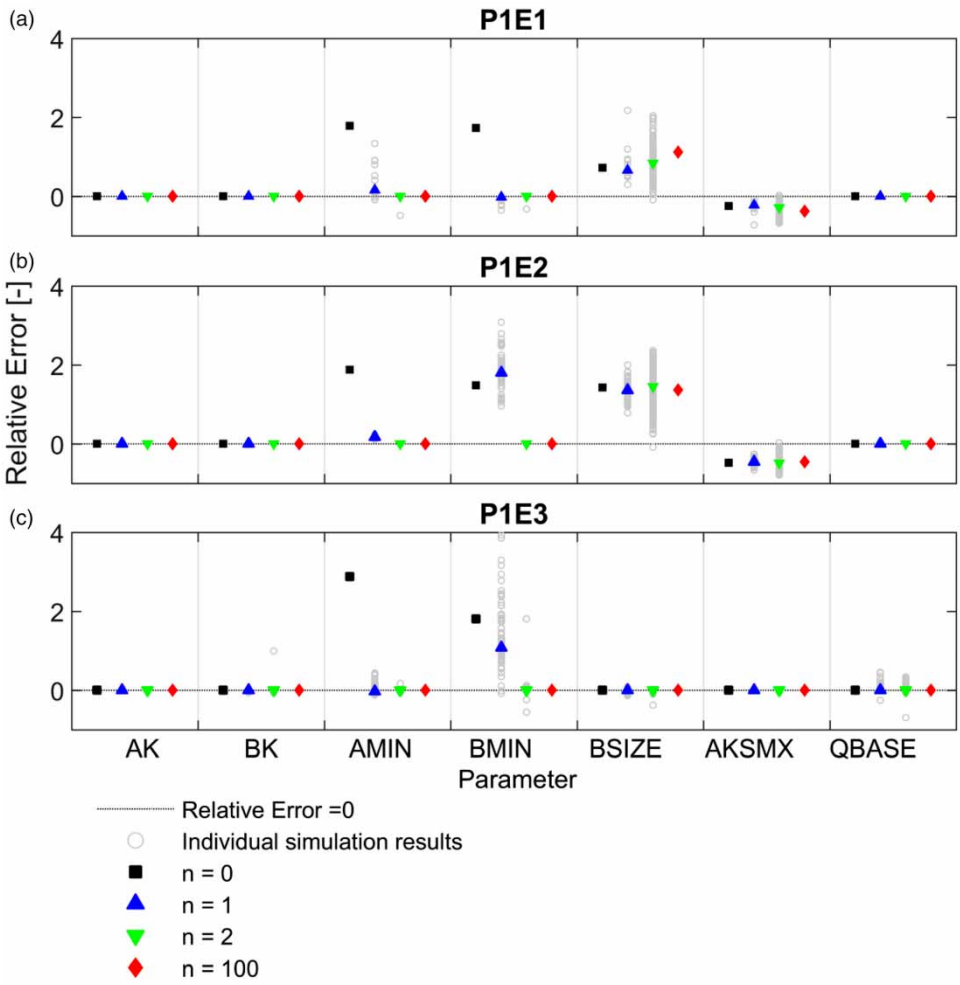

Figure 8 | Median relative error (the difference between the calibrated and real parameter value divided by the real parameter value) for the seven parameters when the model was calibrated with no (square), 1 (triangle), 2 (triangle) and 100 (diamond) isotope samples for parameterization P1 and the three events (small event top row, medium event middle row, large event bottom row). The open circles represent individual simulation results.

parameter (Figures 3 and 9 and Figure S1). For the slow flow dominated streamflow response, the mean information content of BMIN was higher than for AMIN (Table 3). Because the information content for BMIN was generally highest at the end of the event when the fast response flow (Q_A) had ended, samples taken during this period were considered most informative because they allowed the identification of parameter BMIN (Figures 3 and 9 and Figure S1). For the fast flow dominated streamflow response, the mean information content of AMIN was higher than for BMIN (Table 3). Because the information content for AMIN was highest near peakflow, samples taken near peakflow conditions were considered most informative for these situations (Figures 3 and 9 and Figure S1). When overflow occurred, the mean information content of AMIN decreased but was highest at the start or end of overflow and samples taken at these times were most informative to identify AMIN (Figures 3 and 9 and Figure S1).

DISCUSSION

Number of samples needed for model calibration

The multi-criteria model calibration based on the isotopic composition of stream water reduced the parameter

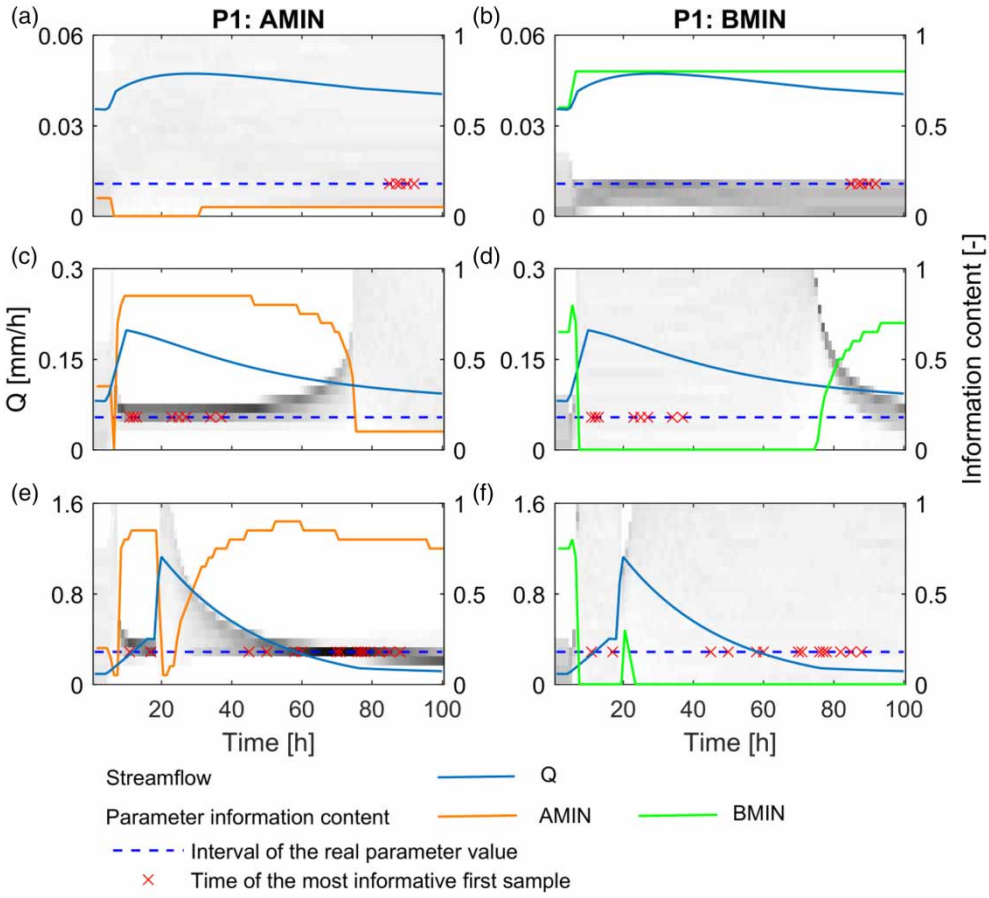

Figure 9 | Information content for parameters AMIN (left column) and BMIN (right column) for parameterization P1 based on the DYNIA analysis, with streamflow and the times of the most informative first samples (crosses) for the small event (top row), medium event (middle row) and large event (bottom row). A darker background colour indicates a higher density of the selected parameters in a certain interval and thus better parameter identifiability. The dashed line represents the interval of the real parameter value. Please see the online pdf of the paper for the color version of this figure.

Table 3 | Mean information content (I) and the standard deviation of the information content (SD) for parameters AMIN and BMIN at the selected most informative first sampling times (see crosses in Figure 3 and Figure S1)

	Event 1				Event 2				Event 3			
	I_AMIN	SD	I_BMIN	SD	I_AMIN	SD	I_BMIN	SD	I_AMIN	SD	I_BMIN	SD
P1	0.05	0	0.80	0	0.85	0	0	0	0.80	0.03	0	0
P2	0.05	0	0.80	0	0.80	0.03	0	0	0.75	0.04	0	0
P3	0.05	0	0.80	0	0.85	0.04	0	0	0.60	0.03	0	0
P4	0.13	0.07	0.80	0.02	0.85	0.03	0	0	0.88	0.05	0	0
P5	0	0.02	0.80	0	0.15	0	0.65	0.03	0.40	0.03	0.33	0.18
P6	0.05	0.02	0.80	0	0.90	0.03	0	0	0.90	0.03	0	0
P7	0.05	0	0.80	0	0.90	0.02	0	0	0.95	0.02	0	0
P8	0	0	0.80	0	0.75	0	0	0	0.85	0.43	0	0.18
P9	0.05	0	0.80	0	0.95	0.02	0	0	0.80	0	0	0

uncertainty for parameters AMIN and BMIN (Figure 8) and resulted in parameter sets that better represented the internal processes, which is consistent with the results of other multi-criteria calibration studies (de Grosbois *et al.* 1988; Seibert 2000; Birkel *et al.* 2010a, 2010b). Similar to other studies (Bergström *et al.* 2002; McGuire *et al.* 2007; Stadnyk *et al.* 2013), the error in the simulated streamflow increased slightly by adding the isotope data but the improvement in the simulation of the isotopic composition of stream water outweighed the decrease in the simulation of streamflow. Surprisingly, our results show that a few isotope samples were sufficient to reduce the parameter uncertainty and improve the internal consistency of the model for the situations in this study, i.e., when there are no errors in the model or the data. Previous studies using isotope data for model calibration used many more samples. For example, Weiler *et al.* (2003) used hourly data for two events for model calibration, while Birkel *et al.* (2010a) used daily data for a one-year period for model calibration. However, the results are in agreement with McIntyre & Wheater (2004), who tested the value of stream phosphorus data for the calibration of a stream phosphorus model and showed that four measurements taken during an event were as informative for model calibration as nine weekly samples and 62 daily samples, also when there were data errors and model structural errors.

The number of samples that can be collected and analysed manually or with automatic samplers is often restricted by practical and financial constraints. The improvement in parameter uncertainty and model consistency based on the small number of samples holds great promise for model calibration for catchments where stream water is currently not regularly sampled because it is more cost-effective to only take a few samples during an event than to obtain daily or weekly samples for a longer period. The fact that the exact timing of the samples is not so important when more than one sample is available for model calibration further reduces the logistical efforts for sampling and suggests that it will be beneficial to take a stream water sample when gauging stations are visited.

Best time for sample collection

Sampling the rising limb or at peak streamflow is challenging in fast responding catchments with very short response times. The results of this study suggest that these samples are less informative for model calibration than samples taken on the falling limb. Sampling late in the event is logistically much easier than sampling at peakflow or during the rising limb due to the longer lead time for getting to the sampling location. In fact, the samples that were considered most informative for model calibration for the slow response and fast response with overflow dominated situations were mostly on the falling limb and often after what would be considered the end of the event (Figure 3 and Figure S1). Many hydrologists would not have bothered to take samples this late during an event, but this study shows that such samples are very informative for model calibration when only a few samples (in this case only one sample) are available. Even for the fast response dominated systems without overflow, the most informative sample was often just after peak streamflow.

The results of the study also suggest that when more than one sample is taken during the event, the exact timing of the sample is not that important, as for most of the events and parameterizations the calibration based on two or three random samples led to similar maximum errors in the simulated isotopic composition of stream water as for the intelligently selected samples. For only two of the 27 runoff events were more than three samples (i.e., six for P8E3 and four for P9E3) needed.

However, it should be noted that the late timing of the most informative isotope samples for model calibration and the small number of samples required for model calibration are very different from the data requirements for other studies, such as hydrograph separation, transit time estimation and load estimation, for which samples on the rising limb and at peakflow are very important (Thomas & Lewis 1993; Littlewood 1995; Robertson & Roerish 1999; Macrae *et al.* 2007; Duvert *et al.* 2011; Hrachowitz *et al.* 2011).

The DYNIA results suggest that the selected most informative sampling times correspond to the periods with the highest information content for certain parameters. Even though we maintained the values of the other parameters at their real value, which is not possible for real catchments because the parameter values would be unknown, the approach suggests that DYNIA or other parameter identifiability analyses are very useful for providing guidance on sampling strategies to improve model calibration. If initial model calibration shows that a certain model parameter

has a large uncertainty, then the parameter identifiability test can provide guidance on when to take samples to reduce parameter uncertainty and improve model consistency. Several researchers have commented on the need for iterative model development, where field data guides model development and model results guide further field measurements, which should then lead to further model improvement (Son & Sivapalan 2007; Fenicia *et al.* 2008; Hrachowitz *et al.* 2014). The use of parameter identifiability analysis to determine when to take stream water samples to improve model calibration appears to be a suitable way to do this, particularly when initial model simulations are combined with information on the expected size of the event during which samples will be taken.

Limitations of this virtual study and applicability to the real world

The main results were similar for a range of different types of catchment responses (as represented by the different parameterizations), including very slow runoff responses (e.g., P5) and very quick runoff responses (e.g., P9). Initial tests, furthermore, suggested that doubling the rainfall intensity did not affect the number of samples required for model calibration or the most informative sampling time. This suggests that the results of this study are applicable for a wide range of situations. However, we used synthetic data to simulate streamflow and the isotopic composition of stream water for a single rainfall–runoff event, rather than real data or a series of events. Changes in the isotopic signal of the rainfall during the event were not considered either.

We simulated a single rainfall event in order to determine the most informative sampling time during an event, rather than the most informative type of event or antecedent conditions. However, we expect that the inclusion of the information from two stream water samples will also cause a better model fit when the model is applied to multiple rainfall events because the inclusion of the isotope data resulted in a reduction in parameter uncertainty. Several previous studies have demonstrated that event-based sampling (particularly during a large event) provides valuable information for model calibration and streamflow simulation for longer periods (McIntyre & Wheater 2004; Juston *et al.* 2009; Seibert & McDonnell 2015).

The use of synthetic data allowed us to obtain a perfect model fit, to have continuous stream isotope data and to obtain clear patterns in the effect of the sampling time on model calibration. This is not the case for real catchments where the model structure does not capture all hydrological processes and a perfect fit cannot be obtained. While complete mixing does not occur in real aquifers or catchments, complete mixing is often considered a useful approximation and has the advantage of not requiring any additional parameters. The response that we see in the stream often looks similar to complete mixing because of the mixing of water from different parts of the catchment. Regardless, the effects of different model structures on the timing of the most informative sample require further research.

Real data are influenced by measurement uncertainties and may be dis-informative, which also limits how well a model can fit the data (Beven & Westerberg 2011; McMillan *et al.* 2012; Beven 2015). For the synthetic data used in this study, the streamflow data already contained sufficient information to constrain five of the seven parameters and the isotope data was needed only to constrain the two other parameters. We expect that when a perfect model fit for streamflow cannot be obtained, the parameter uncertainty for these five parameters will be larger and that additional samples may help to reduce parameter uncertainty for some of these parameters as well. This would mean that (a small number of) additional samples would help to improve model calibration and thus more than two samples are needed for calibration. However, McIntyre & Wheater (2004) showed that errors in the data and model structure limited the value of calibration data severely and that model performance deteriorated, despite reasonable performance for the calibration conditions. In this case, more samples might not add more information for model calibration. The effects of measurement errors on the number of samples for model calibration and the best times for sampling, therefore, needs to be studied further.

The choice of the objective function and optimization algorithm for model calibration might also have affected the results, as shown in other studies (Moussa & Chahinian 2009; Jie *et al.* 2015). However, the use of synthetic data allowed us to obtain model fits for streamflow that were almost perfect (very small values of the objective function for streamflow (F_Q); Table 2). Therefore, the calibration

results are not likely significantly influenced by the choice of objective function for streamflow (Equation (2)). Similarly, the values of the objective function for the isotopic composition of stream water were also very small when two or more stream water samples were used (Table 2). Therefore, we assume that these choices did not significantly affect the results for the best sampling times.

CONCLUSION

Using synthetic data for nine parameterizations and three different rainfall events, we showed that only a few isotope samples are needed to reduce model parameter uncertainty and improve internal model consistency. When only one sample was available, the sampling time influenced model calibration. Intelligently selected samples performed better than other benchmark selections with lower values for the objective function and smaller parameter ranges. Surprisingly, in most cases, a sample taken on the falling limb of the event was most informative for model calibration and was more informative than a sample taken on the rising limb. For slow flow dominated responses and fast flow dominated responses with overflow, the most informative samples for model calibration were often near the end of the event; for fast flow dominated responses without overflow, the most informative samples were near or after peak flow. The times of the most informative samples for model calibration corresponded to the times with the highest information content for the two parameters that could not be determined based on streamflow data alone (AMIN and BMIN, the threshold storage for the fast and slow response flow to occur, respectively). The sampling time did not influence the calibration when two or more samples were available, except for the large rainfall events for P8 and P9. In short, a few selected samples can be very useful for model calibration, and the timing of the most informative sample depends on the flow response but is often on the falling limb of the hydrograph. The results, furthermore, suggest that parameter identifiability analysis can provide information on when to take water quality samples to reduce parameter uncertainty and improve model consistency, which may be useful for iterative model calibration in the real world. Overall, these results provide guidance for cost-effective sampling for model calibration but need to be confirmed with real data, and tested with different coupled flow and tracer models.

ACKNOWLEDGEMENTS

We thank Sergio Maffioletti for IT support for ScienceCloud at the University of Zurich, which enabled us to run the computational-intensive simulations on virtual machines. We thank Sandra Pool, Benjamin Fischer and Marc Vis for helpful discussions and the reviewers for their suggestions to improve this manuscript. This work was funded by the Swiss National Science Foundation (Project-143995).

REFERENCES

Bergström, S., Lindström, G. & Pettersson, A. 2002 Multi-variable parameter estimation to increase confidence in hydrological modelling. *Hydrol. Process.* **16**, 413–421. doi:10.1002/hyp.332.

Beven, K. 2015 Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrol. Sci. J.* **61** (9), 1652–1655. doi:10.1080/02626667.2015.1031761.

Beven, K. & Westerberg, I. 2011 On red herrings and real herrings: disinformation and information in hydrological inference. *Hydrol. Process.* **25**, 1676–1680. doi:10.1002/hyp.7963.

Birkel, C. & Soulsby, C. 2015 Advancing tracer-aided rainfall-runoff modelling: a review of progress, problems and unrealised potential. *Hydrol. Process.* **29**, 5227–5240. doi:10.1002/hyp.10594.

Birkel, C., Dunn, S. M., Tetzlaff, D. & Soulsby, C. 2010a Assessing the value of high-resolution isotope tracer data in the stepwise development of a lumped conceptual rainfall-runoff model. *Hydrol. Process.* **24**, 2335–2348. doi:10.1002/hyp.7763.

Birkel, C., Tetzlaff, D., Dunn, S. M. & Soulsby, C. 2010b Towards a simple dynamic process conceptualization in rainfall-runoff models using multi-criteria calibration and tracers in temperate, upland catchments. *Hydrol. Process.* **24**, 260–275. doi:10.1002/hyp.7478.

Birkel, C., Tetzlaff, D., Dunn, S. M. & Soulsby, C. 2011 Using lumped conceptual rainfall-runoff models to simulate daily isotope variability with fractionation in a nested mesoscale catchment. *Adv. Water Resour.* **34**, 383–394. doi:10.1016/j.advwatres.2010.12.006.

Christophersen, N. & Wright, R. F. 1981 Sulfate budget and a model for sulfate concentrations in stream water at Birkenes,

a small forested catchment in southernmost Norway. *Water Resour. Res.* **17**, 377–389. doi:10.1029/WR017i002p00377.

de Grosbois, E., Dillon, P. J., Seip, H. M. & Seip, R. 1986 Modelling hydrology and sulphate concentration in small catchments in Central Ontario. *Water Air Soil Pollut.* **31**, 45–57. doi:10.1007/BF00630818.

de Grosbois, E., Hooper, R. P. & Christophersen, N. 1988 A multisignal automatic calibration methodology for hydrochemical models: a case study of the Birkenes Model. *Water Resour. Res.* **24**, 1299–1307. doi:10.1029/WR024i008p01299.

Duan, Q., Sorooshian, S. & Gupta, V. 1992 Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **28**, 1015–1031. doi:10.1029/91WR02985.

Duan, Q. Y., Gupta, V. K. & Sorooshian, S. 1993 Shuffled complex evolution approach for effective and efficient global minimization. *J. Optim. Theory Appl.* **76**, 501–521. doi:10.1007/BF00939380.

Duan, Q., Sorooshian, S. & Gupta, V. K. 1994 Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.* **158**, 265–284. doi:10.1016/0022-1694(94)90057-4.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. & Wood, E. F. 2006 Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320**, 3–17. doi:10.1016/j.jhydrol.2005.07.031.

Dunn, S. M. & Bacon, J. R. 2008 Assessing the value of Cl – and δ 18 O data in modelling the hydrological behaviour of a small upland catchment in northeast Scotland. *Hydrol. Res.* **39**, 337. doi:10.2166/nh.2008.134.

Duvert, C., Gratiot, N., Némery, J., Burgos, A. & Navratil, O. 2011 Sub-daily variability of suspended sediment fluxes in small mountainous catchments – implications for community-based river monitoring. *Hydrol. Earth Syst. Sci.* **15**, 703–713. doi:10.5194/hess-15-703-2011.

Fenicia, F., McDonnell, J. J. & Savenije, H. H. G. 2008 Learning from model improvement: on the contribution of complementary data to process understanding. *Water Resour. Res.* **44**, 1–13. doi:10.1029/2007WR006386.

Francés, F., Vélez, J. I. & Vélez, J. J. 2007 Split-parameter structure for the automatic calibration of distributed hydrological models. *J. Hydrol.* **332**, 226–240. doi:10.1016/j.jhydrol.2006.06.032.

Grip, H., Jansson, P. E., Johnsson, H. & Nilsson, S. I. 1985 Application of the 'Birkenes' model to two forested catchments on the Swedish west coast. *Ecol. Bull.* **37**, 176–192.

Hooper, R. P., Stone, A., Christophersen, N., de Grosbois, E. & Seip, H. M. 1988 Assessing the Birkenes Model of stream acidification using a multisignal calibration methodology. *Water Resour. Res.* **24**, 1308–1316. doi:10.1029/WR024i008p01308.

Hrachowitz, M., Soulsby, C., Tetzlaff, D. & Malcolm, I. A. 2011 Sensitivity of mean transit time estimates to model conditioning and data availability. *Hydrol. Process.* **25**, 980–990. doi:10.1002/hyp.7922.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G. & Gascuel-Odoux, C. 2014 Process consistency in models: the importance of system signatures, expert knowledge, and process complexity. *Water Resour. Res.* **50**, 7445–7469. doi:10.1002/2014WR015484.

Hrachowitz, M., Fovet, O., Ruiz, L. & Savenije, H. H. G. 2015 Transit time distributions, legacy contamination and variability in biogeochemical 1/f α scaling: how are hydrological response dynamics linked to water quality at the catchment scale? *Hydrol. Process.* **29**, 5241–5256. doi:10.1002/hyp.10546.

Jie, M., Chen, H., Xu, C., Zeng, Q. & Tao, X. 2015 A comparative study of different objective functions to improve the flood forecasting accuracy. *Hydrol. Res.* **2**, nh2015078. doi:10.2166/nh.2015.078.

Juston, J., Seibert, J. & Johansson, P.-O. 2009 Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. *Hydrol. Process.* **23**, 3093–3109. doi:10.1002/hyp.7421.

Kirchner, J. W. 2003 A double paradox in catchment hydrology and geochemistry. *Hydrol. Process.* **17**, 871–874. doi:10.1002/hyp.5108.

Kirchner, J. W. 2006 Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res.* **42**, 1–5. doi:10.1029/2005WR004362.

Leibundgut, C., Maloszewski, P. & Klls, C. 2009 *Tracers in Hydrology*, 1st edn. John Wiley & Sons, Ltd, Chichester, UK. doi:10.1002/9780470747148.

Lindström, G. 2016 Lake water levels for calibration of the S-HYPE model. *Hydrol. Res.* **47**, 672–682. doi:10.2166/nh.2016.019.

Lindström, G. & Rodhe, A. 1986 Modelling water exchange and transit times in till basins using oxygen-18. *Hydrol. Res.* **17**, 325–334.

Littlewood, I. 1995 Hydrological regimes, sampling strategies, and assessment of errors in mass load estimates for United Kingdom rivers. *Environ. Int.* **21**, 211–220. doi:10.1016/0160-4120(95)00011-9.

Macrae, M. L., English, M. C., Schiff, S. L. & Stone, M. 2007 Capturing temporal variability for estimates of annual hydrochemical export from a first-order agricultural catchment in southern Ontario, Canada. *Hydrol. Process.* **21**, 1651–1663. doi:10.1002/hyp.6361.

McGuire, K. J., Weiler, M. & McDonnell, J. J. 2007 Integrating tracer experiments with modeling to assess runoff processes and water transit times. *Adv. Water Resour.* **30**, 824–837. doi:10.1016/j.advwatres.2006.07.004.

McIntyre, N. R. & Wheater, H. S. 2004 Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty. *J. Hydrol.* **290**, 100–116. doi:10.1016/j.jhydrol.2003.12.003.

McMillan, H., Krueger, T. & Freer, J. 2012 Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrol. Process.* **26**, 4078–4111. doi:10.1002/hyp.9384.

Moussa, R. & Chahinian, N. 2009 Comparison of different multi-objective calibration criteria using a conceptual rainfall-runoff model of flood events. *Hydrol. Earth Syst. Sci.* **13**, 519–535. doi:10.5194/hess-13-519-2009.

Neal, C., Christophersen, N., Neale, R., Smith, C. J., Whitehead, P. G. & Reynolds, B. 1988 Chloride in precipitation and streamwater for the upland catchment of River Severn, mid-Wales; some consequences for hydrochemical models. *Hydrol. Process.* **2**, 155–165. doi:10.1002/hyp.3360020206.

Page, T., Beven, K. J., Freer, J. & Neal, C. 2007 Modelling the chloride signal at Plynlimon, Wales, using a modified dynamic TOPMODEL incorporating conservative chemical mixing (with uncertainty). *Hydrol. Process.* **21**, 292–307. doi:10.1002/hyp.6186.

Perrin, C., Oudin, L., Andereassian, V., Rojas-Serna, C., Michel, C. & Mathevet, T. 2007 Impact of limited streamflow data on the efficiency and the parameters of rainfall–runoff models. *Hydrol. Sci. J.* **52**, 131–151. doi:10.1623/hysj.52.1.131.

Pianosi, F., Sarrazin, F. & Wagener, T. 2015 A Matlab toolbox for global sensitivity analysis. *Environ. Model. Softw.* **70**, 80–85. doi:10.1016/j.envsoft.2015.04.009.

Raat, K. J., Vrugt, J. A., Bouten, W. & Tietema, A. 2004 Towards reduced uncertainty in catchment nitrogen modelling: quantifying the effect of field observation uncertainty on model calibration. *Hydrol. Earth Syst. Sci.* **8**, 751–763. doi:10.5194/hess-8-751-2004.

Robertson, D. M. & Roerish, E. D. 1999 Influence of various water quality sampling strategies on load estimates for small streams. *Water Resour. Res.* **35**, 3747–3759. doi:10.1029/1999WR900277.

Rustad, S., Christophersen, N., Seip, H. M. & Dillon, P. J. 1986 Model for streamwater chemistry of a tributary to Harp Lake, Ontario. *Can. J. Fish. Aquat. Sci.* **43**, 625–633.

Seibert, J. 2000 Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrol. Earth Syst. Sci.* **4**, 215–224. doi:10.5194/hess-4-215-2000.

Seibert, J. & McDonnell, J. J. 2015 Gauging the ungauged basin: relative value of soft and hard data. *J. Hydrol. Eng.* **20**, A4014004. doi:10.1061/(ASCE)HE.1943-5584.0000861.

Seibert, J., Rodhe, A. & Bishop, K. 2003 Simulating interactions between saturated and unsaturated storage in a conceptual runoff model. *Hydrol. Process.* **17**, 379–390. doi:10.1002/hyp.1130.

Seip, H. M., Seip, R., Dillon, P. J. & de Grosbois, E. 1985 Model of sulphate concentration in a small stream in the Harp Lake catchment, Ontario. *Can. J. Fish. Aquat. Sci.* **42**, 927–937.

Son, K. & Sivapalan, M. 2007 Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resour. Res.* **43**, 1–18. doi:10.1029/2006WR005032.

Soulsby, C., Birkel, C., Geris, J., Dick, J., Tunaley, C. & Tetzlaff, D. 2015 Stream water age distributions controlled by storage dynamics and nonlinear hydrologic connectivity: modeling with high-resolution isotope data. *Water Resour. Res.* **51**, 7759–7776. doi: 10.1002/2015WR017888.

Stadnyk, T. A., Delavau, C., Kouwen, N. & Edwards, T. W. D. 2013 Towards hydrological model calibration and validation: simulation of stable water isotopes using the isoWATFLOOD model. *Hydrol. Process.* **3810**, 3791–3810. doi:10.1002/hyp.9695.

Thomas, R. B. & Lewis, J. 1993 A comparison of selection at list time and time-stratified sampling for estimating suspended sediment loads. *Water Resour. Res.* **29**, 1247–1256. doi:10.1029/92WR02711.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S. & Gupta, H. V. 2003 Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. *Hydrol. Process.* **17**, 455–476. doi:10.1002/hyp.1135.

Weiler, M., McGlynn, B. L., McGuire, K. J. & McDonnell, J. J. 2003 How does rainfall become runoff? A combined tracer and runoff transfer function approach. *Water Resour. Res.* **39**. doi:10.1029/2003WR002331.

Wheater, H. S., Bishop, K. H. & Beck, M. B. 1986 The identification of conceptual hydrological models for surface water acidification. *Hydrol. Process.* **1**, 89–109. doi:10.1002/hyp.3360010109.

Yapo, P. O., Gupta, H. V. & Sorooshian, S. 1996 Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. *J. Hydrol.* **181**, 23–48. doi:10.1016/0022-1694(95)02918-4.

First received 21 July 2016; accepted in revised form 23 November 2016. Available online 13 March 2017